I-Arabic: Computational Attempts and Corpus Issues in Modern Arabic

العربية في عصر التكنولوجيا: مقاربات حاسوبية وإشكاليات المتون في اللغة العربية الحديثة

Nagwa Younis *
nagwayounis@edu.asu.edu.eg

Abstract

Modern Arabic encounters many challenges concerning the use of computer-based methods for analyzing Arabic data. These methods include natural language processing, machine learning, and corpus linguistics, among others. This paper addresses the challenges, the computational attempts, and a proposed model: I-Arabic. One of the main challenges in using computational methods for Arabic is the lack of large, high-quality language resources, such as text corpora, annotated data, and lexical resources. This is due to various factors, including the diversity of Arabic dialects and the limited availability of digitized Arabic texts. Another challenge is the complexity of Arabic morphology and syntax, which can pose difficulties for natural language processing algorithms. Arabic is a highly inflected language, with a rich system of prefixes, suffixes, and internal vowel changes that can affect the meaning and function of words. Additionally, Arabic has a flexible word order and a complex system of grammatical agreement. Despite these challenges, there have been significant efforts to develop computational tools and resources for Arabic, including the creation of Arabic language corpora and the development of natural language processing algorithms specifically tailored to Arabic. These efforts have the potential to facilitate research in various fields, including linguistics, social media analysis, and machine translation.

Keywords: Modern Arabic, Computational Attempts, Arabic Corpus, I-Arabic.

* **Professor of Linguistics, AOU.**

الملخَّص:

تواجه اللغة العربية الحديثة العديد من التحديات المتعلقة باستخدام الأساليب القائمة على الكمبيوتر لتحليل البيانات العربية. تشمل هذه الأساليب معالجة اللغة الطبيعية، والتعلم الآلي، ولغويات المتن، من بين أمور أخرى. تتناول هذه الورقة البحثية التحديات والمحاولات الحاسوبية والنموذج المقترح: I-Arabic.. أحد التحديات الرئيسية في استخدام الأساليب الحاسوبية للغة العربية هو الافتقار إلى موارد لغوية كبيرة وعالية الجودة، مثل: متن النصوص، والبيانات المشروحة، والموارد المعجمية. ويرجع ذلك إلى عوامل مختلفة، بما في ذلك تنوع اللهجات العربية، ومحدودية توافر النصوص العربية الرقمية. والتحدي الآخر هو تعقيد الصرف العربي وبناء الجملة، والذي يمكن أن يشكل صعوبات لخوارزميات معالجة اللغة الطبيعية. تعتبر اللغة العربية لغةً شديدة التصريف، مع نظام غني من السوابق واللواحق، وتغييرات حروف العلة الداخلية التي يمكن أن تؤثر على معنى الكلمات ووظيفتها. بالإضافة إلى ذلك، تتمتع اللغة العربية بترتيب كلمات مرن، ونظام معقد من الاتفاق النحوي. وعلى الرغم من هذه التحديات، كانت هناك جهود كبيرة لتطوير الأدوات والموارد الحاسوبية للغة العربية، بما في ذلك إنشاء مجموعة اللغة العربية، وتطوير خوارزميات معالجة اللغة الطبيعية المصممة خصيصا للغة العربية. هذه الجهود لديها القدرة على تسهيل البحث في مختلف المجالات، بما في ذلك اللغويات وتحليل وسائل التواصل الاجتماعي والترجمة الآلية.

الكلمات المفتاحية: اللغة العربية الحديثة، محاولات حاسوبية، المتون العربية، I-Arabic.

1. Introduction

In contradistinction with other modern world languages, Modern Arabic is a complex language with a rich history and diverse dialects. The Arabic language is one of the most widely spoken languages in the world, with over 420 million speakers across the Middle East and North Africa, as well as in diaspora communities around the world. Arabic is the official language of 27 countries, including Egypt, Saudi Arabia, and Iraq, and is one of the six official languages of the United Nations. Despite its widespread use, the status of the Arabic language today is complex and multifaceted. There are some key aspects of the current status of Arabic. They include:

1. **Standard Arabic vs. Dialects**: Standard Arabic, which is based on the classical Arabic of the Holy Qur'an, is the official language of most Arabic-speaking countries. However, many people in the region speak dialects of Arabic that are different from Standard Arabic. While Standard Arabic is used in formal settings like government, media, and education, dialects are more commonly used in everyday conversation.

2. **Language Policy**: In many Arabic-speaking countries, there are language policies in place that promote the use of Standard Arabic and discourage the use of dialects in official contexts. However, there is also recognition of the importance of dialects as a part of the region's linguistic heritage.

3. **Arabic Education**: Arabic is taught in schools throughout the Middle East and North Africa (MENA region), and is often a mandatory subject. However, the quality of Arabic education varies widely, and there are concerns about the effectiveness of teaching methods and the relevance of the curriculum.

4. **Technology**: The use of modern technology has had a significant impact on the Arabic language. While technology has made Arabic content more accessible than ever before, there are also concerns about the impact of technology on the quality of Arabic language use, particularly in informal contexts like social media.

The status of the Arabic language today is shaped by a complex interplay of linguistic, cultural, and political factors. While Arabic remains an important language in the region and beyond, there are also challenges to its continued use and vitality. Hence, Computational attempts are needed to address the challenges of working with Arabic corpora, including issues related to morphology, orthography, and

dialectal variation. One of the main challenges in working with Arabic corpora is the complexity of Arabic morphology. Arabic words can have multiple forms depending on their context and grammatical function, and these forms can include prefixes, suffixes, and infixes. To address this challenge, computational attempts such as morphological analyzers and stemmers have been developed. These tools can analyze Arabic words and provide information about their root form, grammatical gender, and other features. Another challenge in working with Arabic corpora is the issue of orthography. Arabic script is written from right to left, and Arabic letters can change shape depending on their position in a word. This can make it difficult to tokenize Arabic text and identify individual words. To address this challenge, computational solutions such as tokenizers and part-of-speech taggers have been developed. These tools can identify individual words and provide information about their grammatical function (McEnery et al., 2019).

Dialectal variation is also a major challenge in working with Arabic corpora, as Arabic is spoken in many different dialects across the Middle East and North Africa (MENA region). These dialects can differ significantly in terms of vocabulary, grammar, and pronunciation. To address this challenge, computational solutions such as dialect classifiers and language models have been developed. These tools can identify the dialect of a given text and provide information about dialect-specific features. To this end, computational attempts are crucial for addressing the challenges of working with Arabic corpora. These solutions can help researchers and language professionals to analyze and understand Arabic language data more effectively, and to develop more accurate and comprehensive language models for Arabic (Habash, 2010).

2. Challenges for Developing Modern Arabic Computational Models

Developing language models for Arabic presents a number of challenges due to the complexity and diversity of the language. Some of the main challenges include:

1. **Morphological complexity**: As mentioned earlier, Arabic has a complex morphology with words having multiple forms. This poses a

challenge for developing accurate language models that can handle the various forms of words.

2. **Dialectal variation**: Arabic is spoken in many different dialects across the Middle East and North Africa (MENA region), each with its own unique vocabulary, grammar, and pronunciation. This makes it challenging to develop language models that can accurately capture the nuances of each dialect.

3. **Limited resources**: Compared to other languages, Arabic has limited digital resources such as annotated corpora, lexicons, and speech databases. This makes it difficult to train and evaluate language models effectively.

4. **Orthographic variation**: Arabic script is written from right to left, and letters can take on different shapes depending on their position in a word. This can make it challenging to develop language models that can accurately identify and parse individual words.

5. **Ambiguity**: Arabic words can have multiple meanings depending on their context, which makes it challenging to develop language models that can accurately disambiguate words.

6. **Script conversion**: Arabic is often transliterated into other scripts for use in digital applications. This can introduce errors and inconsistencies that make it challenging to develop accurate language models.

3. Computational Attempts for Modelling Modern Arabic

These challenges require innovative solutions and ongoing research to improve the accuracy and effectiveness of language models for Arabic. There are a variety of computational Arabic applications that have been developed in recent years (e.g. Belinkov et al., 2013) . The following are some examples:

1. **Machine translation**: Machine translation technology can be used to translate text from Arabic to other languages and vice versa. Machine translation can be particularly helpful for communication between speakers of different languages, such as in international business or diplomacy. There are several Arabic machine translation systems available today, including *Google Translate*, *Microsoft Translator*, and *Systran*.

2. **Speech recognition**: Speech recognition technology can be used to transcribe spoken Arabic into written text, making it easier to store, search, and share. This technology can be particularly helpful for individuals with disabilities who may have difficulty typing or writing. It is used in applications such as virtual assistants and automated customer service systems.

4. **Text-to-speech**: Text-to-speech technology can be used to convert written Arabic text into spoken language, making it easier for individuals with visual impairments to access written information. This technology is used in applications such as automated call centers, audiobooks, and language learning software.

4. **Language learning software**: Language learning software can be used to teach Arabic language learners the basics of the language, including grammar, vocabulary, and pronunciation. This technology can be particularly helpful for individuals who are unable to attend traditional language classes.

5. **Sentiment Analysis**: Sentiment analysis is the process of identifying the sentiment or emotion conveyed by a piece of text. This technique is used in Arabic social media analysis, customer feedback analysis, and market research.

6. **Named Entity Recognition**: Named entity recognition is the process of identifying and classifying named entities such as people, organizations, and locations in a piece of text. This technique is used in Arabic information extraction, news analysis, and search engines.

4. Arabic Optical Character Recognition (OCR)

Arabic OCR technology is used to recognize printed or handwritten Arabic text and convert it into digital text. This technology is used in applications such as document digitization, text recognition, and content analysis. OCR technology can be particularly useful for preserving historical documents and manuscripts, which may be written in Arabic script. OCR technology can be used to preserve Arabic manuscripts by digitizing them and making them easier to store, search, and share. Here are some ways in which OCR technology can be used to preserve Arabic manuscripts (Alghyaline, 2023). Several privileges can be attributed to :

1. **Digitization:**

OCR technology can be used to digitize printed Arabic manuscripts, converting them into digital format. This makes it easier to store them in digital archives and libraries, where they can be accessed by researchers and scholars from around the world.

2. **Searchability:**

Once Arabic manuscripts have been digitized using OCR technology, they can be made searchable using text recognition algorithms. This makes it easier to find specific words, phrases, or topics within the manuscripts, saving researchers time and effort.

3. **Preservation:**

Digitizing Arabic manuscripts using OCR technology can help to preserve them by reducing the need for physical handling and exposure to environmental factors. Digital copies of Arabic manuscripts can be stored in secure, climate-controlled archives, helping to prevent damage and decay over time.

4. **Accessibility:**

Digitized Arabic manuscripts can be made accessible to a wider audience, including individuals with visual impairments who may not be able to read printed manuscripts. By converting printed manuscripts into digital format, OCR technology can facilitate access to these important cultural artifacts for people who might not otherwise be able to experience them.

Moreover, OCR technology has the potential to revolutionize the preservation and accessibility of Arabic manuscripts by converting them into digital format. By digitizing Arabic manuscripts using OCR technology, we can ensure that these important cultural artifacts are preserved for future generations and made accessible to a wider audience of researchers and scholars around the world. Digitizing Arabic manuscripts using OCR technology can be challenging due to the complexity and variability of the Arabic script. However, some challenges can arise when using OCR technology to digitize Arabic manuscripts:

1. **Arabic script variations**: The Arabic script includes many variations of letters and diacritical marks that can be difficult for OCR technology to recognize and distinguish from one another. This can lead to errors in the OCR output, particularly in cases where the script is handwritten or contains unusual or archaic forms.

2. **Historical context**: Arabic manuscripts can be written in a variety of historical scripts and styles that may not be familiar to modern OCR technology. This can make it difficult for OCR software to accurately recognize and transcribe these manuscripts.

3. **Quality of the source material**: The quality of the source material for Arabic manuscripts can vary widely, with some manuscripts being damaged, faded, or poorly preserved. This can make it difficult for OCR technology to accurately recognize and transcribe the text.

4. **Lack of standardization**: There is no standardized orthography for Arabic manuscripts, which can make it difficult for OCR technology to accurately recognize and transcribe the text. This is particularly true for older manuscripts, which may use spellings and diacritical marks that are no longer used in modern Arabic.

5. **Language-specific challenges**: Arabic OCR technology faces some specific challenges such as the presence of non-standard diacritics, the absence of vowel letters in some texts, and the need for special fonts.

5. Digitization Projects for Arabic Manuscripts

Several successful digitization projects for Arabic manuscripts have been attempted in recent years. Some examples of which are:

1. **The Digital Library of the Middle East**: The Digital Library of the Middle East (https://dlmenetwork.org/library) is a collaborative project among several institutions in the Middle East, Europe, and the United States. The project aims to digitize and make available online a wide range of manuscripts, maps, and other cultural artifacts from the Middle East.

2. **The Qatar Digital Library**: The Qatar Digital Library is a project of the Qatar National Library, which aims to digitize and make available online a wide range of manuscripts, maps, and other cultural artifacts from the Middle East and beyond. The project includes a large collection of Arabic manuscripts, many of which have been digitized using OCR technology.

3. **The Digital Scriptorium**: The Digital Scriptorium is a project of several institutions in the United States, which aims to digitize and make available online a wide range of manuscripts from around the

world. The project includes a large collection of Arabic manuscripts, many of which have been digitized using OCR technology.

4. **The British Library Qatar Foundation Partnership**: The British Library Qatar Foundation Partnership (https://www.bl.uk/projects/british-library-qatar-foundation-partnership) is a joint project between the British Library and the Qatar Foundation, which aims to digitize and make available online a wide range of Arabic manuscripts from the British Library's collections. The project includes manuscripts from a variety of historical periods and genres, including Islamic law, science, and literature.

These successful digitization projects demonstrate the potential of OCR technology to preserve and make accessible important cultural artifacts from the Arabic-speaking world. By digitizing Arabic manuscripts using OCR technology, we can ensure that these important cultural artifacts are preserved for future generations and made accessible to a wider audience of researchers and scholars around the world.

6. Language policy and planning for computing Modern Arabic

An important aspect is developing strategies and guidelines for the effective use of Arabic in digital applications and technologies. This includes addressing issues related to standardization, terminology, and linguistic diversity. Some key considerations for language policy and planning in computing Modern Arabic include:

1. **Standardization:**

Standardization is important for ensuring consistency and interoperability in digital applications and technologies. Language policy and planning should focus on promoting the use of standardized Arabic, including standard Arabic script and vocabulary.

2. **Terminology development:**

The development of a standardized terminology is critical for maintaining consistency and accuracy in the use of Arabic in digital applications. This involves developing a comprehensive set of terms and definitions in Arabic for use in computing and related fields.

3. **Localization:**

Localization involves adapting digital applications and technologies to the linguistic and cultural norms of Arabic-speaking communities. This includes translating user interfaces, documentation,

and marketing materials into Arabic and ensuring that they are culturally appropriate.

4. **Linguistic diversity:**

Arabic is spoken in many different dialects across the Middle East and North Africa. Language policy and planning should take into account the diversity of Arabic and ensure that digital applications and technologies are accessible to speakers of different dialects.

5. **Education and training:**

Education and training are critical for promoting the use of Arabic in computing and related fields. Language policy and planning should support the development of Arabic language and computing curricula in schools and universities, as well as training programs for professionals in the field.

Partially, effective language policy and planning for computing Modern Arabic involves collaboration among linguists, language professionals, and technology experts to promote the use of Arabic in digital applications and technologies.

7. Recent Attempts to Standardise Modern Arabic

Standardized Arabic script and vocabulary are developed and promoted by various language organizations and institutions, such as the International Organization for Standardization (ISO) and the Arab League Educational, Cultural and Scientific Organization (ALECSO) (Elkhafaifi, 2021). Here are some examples of standardized Arabic script and vocabulary:

1. Standard Arabic script: The standard Arabic script follows the rules of the Arabic alphabet and is used for writing formal Arabic. It is also used for transcribing Arabic words in digital applications and technologies. The standard Arabic script is widely recognized and used in Arabic-speaking countries.

2. Modern Standard Arabic (MSA): Modern Standard Arabic is a standardized version of Arabic that is used in formal contexts, such as in writing, media, and education. It is based on the classical Arabic language and is used across the Arabic-speaking world.

3. Arabic Technical Terminology: The Arabic Technical Terminology is a standardized vocabulary that includes technical terms and definitions used in various fields, such as information technology, engineering, and medicine. It is developed and maintained by ALECSO

and is used to ensure consistency and accuracy in the use of Arabic technical terms.

4. ISO 639-3 language codes: The ISO 639-3 language codes are a standardized set of codes used to identify languages and dialects. They include codes for various Arabic dialects, such as Egyptian Arabic (arz), Levantine Arabic (apc), and Gulf Arabic (afb).

5. Arabic Keyboard Layout: The Arabic keyboard layout is a standardized keyboard layout used for typing Arabic characters on a computer. It includes the standard Arabic alphabet and is used in most Arabic-speaking countries.

These examples demonstrate the importance of standardization in promoting the use of Arabic in computing and related fields. Standardized Arabic script and vocabulary help to ensure consistency and accuracy in the use of Arabic in digital applications and technologies, and promote the accessibility of these tools to Arabic speakers.

8. Arabic Language Common Framework (ALCF)

It is worth mentioning that the Arabic Language Common Framework (ALCF) is a set of guidelines developed by the Arab League Educational, Cultural and Scientific Organization (ALECSO) to standardize Arabic language teaching and learning across the Arab world. The ALCF is based on the Common European Framework of Reference for Languages (CEFR), which provides a framework for language proficiency levels and learning objectives. The ALCF is intended to promote the teaching of Arabic as a foreign language and to facilitate communication and cooperation among Arabic language educators across the Arab world. The framework outlines six proficiency levels, ranging from A1 (beginner) to C2 (advanced), and provides descriptors for each level that describe the skills and competencies that learners should be able to demonstrate. The ALCF also includes guidelines for teaching Arabic grammar, vocabulary, and pronunciation, as well as recommendations for assessment and evaluation. The framework emphasizes the importance of communicative competence, cultural awareness, and language learning strategies in Arabic language learning.

The ALCF provides a standardized approach to Arabic language teaching and learning that can help to promote consistency and quality

in Arabic language education across the Arab world. By providing clear learning objectives and assessment criteria, the framework can also help to facilitate the recognition of Arabic language proficiency across different contexts and institutions.

9. Arabic Languoid, Doculect and Glossonym

In linguistics, a "languoid" is a term used to describe a language or dialect that is part of a larger language family. Arabic is a languoid within the larger Afro-Asiatic language family, which includes other languages such as Hebrew, Amharic, and Berber. A "doculect" refers to a specific variety or dialect of a language that is documented through linguistic research or written records. For example, Egyptian Arabic, Moroccan Arabic, and Gulf Arabic are all doculects of Arabic, each with their own unique features and characteristics. A "glossonym" is a term used to describe a specific word or term within a language that is used to provide a gloss or translation of a word in another language (Good, 2013). For example, the Arabic glossonym for the English word "computer" is "الحاسوب" (al-ḥāsūb). These terms are important concepts in linguistics and are used to describe and analyze the diversity and complexity of languages, including Arabic. By understanding these concepts, linguists and language professionals can better understand the linguistic features and characteristics of specific varieties of Arabic, and develop more effective language policy and planning initiatives. Glossonyms can be helpful in language translation by providing a standardized term or word that is used to provide a gloss or translation of a word in another language. These terms are often used in dictionaries, glossaries, and machine translation systems to provide accurate translations of words and phrases. The following are some ways that glossonyms can help in language translation:

1. Standardization: Glossonyms provide a standardized term for a specific word or concept, ensuring consistency and accuracy in translation. This helps to avoid confusion and ambiguity in the translation process.

2. Efficiency: Glossonyms can help to speed up the translation process by providing a ready-made translation for commonly used words and phrases. This can be especially helpful in technical or specialized fields, where specific terminology is used.

3. Accuracy: Glossonyms are often developed and maintained by language experts, ensuring that they are accurate and up-to-date. This can help to ensure that translations are as accurate as possible.

4. Consistency: Glossonyms can help to ensure consistency in translation across different texts and contexts. This is important in fields such as legal or medical translation, where accuracy and consistency are critical.

Hence, glossonyms can be an important tool in language translation, helping to improve accuracy, consistency, and efficiency in the translation process. By providing standardized translations for specific words and phrases, glossonyms can help to bridge linguistic and cultural barriers, facilitating communication and understanding between different languages and cultures.

10.Arabizi and Modern Arabic

Arabizi is a term used to describe the use of the Latin script to write Arabic words, phrases or sentences. It is often used in informal online communication, such as social media, text messaging, and chat rooms, particularly among young Arabs. Arabizi is used as a way to write Arabic words using the Latin script, which is easier to type on a standard keyboard and to read for those who are not familiar with the Arabic script. However, Arabizi is considered by some to be a form of code-switching, as it combines elements of both Arabic and English, and it can be difficult for some people to understand if they are not familiar with the conventions of Arabizi. Moreover, the use of Arabizi can sometimes lead to ambiguity and misunderstandings, as certain Arabic sounds and letters do not have an exact equivalent in the Latin script.

There have been some informal efforts to standardize Arabizi spelling and grammar, particularly in the context of social media and online communication. However, these efforts have not been widely adopted or recognized by official language institutions. Some organizations and individuals have proposed guidelines for Arabizi spelling and grammar, such as the use of certain letters and symbols to represent Arabic sounds and diacritical marks, and the avoidance of English loanwords and expressions.

However, it is important to note that these standardization efforts are still in the early stages and are not yet widely accepted or adopted.

Moreover, there is ongoing debate among linguists and language experts about the usefulness and appropriateness of standardizing Arabizi, given its informal and non-standard nature. There are several challenges in standardizing Arabizi, including:

1. Lack of consensus: There is currently no consensus among experts on how to standardize Arabizi spelling and grammar. There are many different proposals and guidelines, each with their own strengths and weaknesses.

2. Variability: Arabizi is a highly variable and dynamic system, with many different variations and conventions used by different speakers and in different contexts. This makes it difficult to establish a single standard that can be widely adopted.

3. Mixing of languages: Arabizi often involves mixing Arabic and English words and phrases, which can make it difficult to establish consistent spelling and grammar rules.

4. Lack of official recognition: Arabizi is not recognized as an official language or writing system by any language institutions, which makes it difficult to establish a widely accepted standard.

5. Limited use: While Arabizi is used widely in informal online communication, its use in formal contexts such as education and media is limited. This reduces the incentive for standardization efforts and makes it difficult to establish a widely accepted standard.

6. Technical challenges: There are also technical challenges involved in standardizing Arabizi, such as developing software tools and platforms that can support a standardized system, and ensuring compatibility with existing Arabic language technologies and resources.

11.An Academic Attempt: Buckwalter Arabic Morphological Analyzer

It is a tool used for the analysis of Arabic text. Named after its creator, Tim Buckwalter, it is widely used in natural language processing applications such as machine translation, information retrieval, and text-to-speech systems. The Buckwalter system uses a transliteration scheme that maps Arabic characters to ASCII characters, allowing for easy input and processing of Arabic text in computer systems. The scheme assigns a unique ASCII character to each Arabic letter, and also includes symbols for diacritical marks and other non-

letter characters used in Arabic script. Moreover, it uses this transliteration scheme to analyze Arabic text and identify the root form and morphological features of each word. This information can be used for a variety of natural language processing tasks, such as part-of-speech tagging, named entity recognition, and text classification. The Buckwalter Arabic Morphological Analyzer is a powerful tool for natural language processing of Arabic text, and it can be used for a variety of tasks such as:

1. Lemmatization: The analyzer can be used to generate the root form of an Arabic word, which is useful for tasks such as lemmatization (reducing a word to its base form) and stemming (reducing a word to a common stem).

2. Part-of-speech tagging: The analyzer can also be used to identify the part of speech of each word in a sentence, such as noun, verb, adjective, and so on. This information is useful for many natural language processing tasks, such as information retrieval, sentiment analysis, and machine translation.

3. Named entity recognition: The analyzer can help to identify named entities in Arabic text, such as names of people, places, and organizations. This is important for tasks such as information extraction and text classification.

4. Machine translation: The analyzer can be used as a pre-processing step for machine translation systems that translate Arabic text into other languages.

5. Information retrieval: The analyzer can be used to improve the accuracy of information retrieval systems that search for relevant documents or passages in Arabic text.

6. Text normalization: The analyzer can help to normalize Arabic text by converting different forms of the same word to a common base form. This is useful for tasks such as text-to-speech synthesis and information retrieval.

The Buckwalter Arabic Morphological Analyzer is one of the most widely used tools for natural language processing of Arabic text, and it has several advantages over other Arabic NLP tools:

1. Accuracy: The Buckwalter analyzer is known for its high accuracy in identifying the root form and morphological features of Arabic words. This is due to its comprehensive rules and algorithms for Arabic morphology.

2. Speed: The Buckwalter analyzer is relatively fast and efficient, compared to other Arabic NLP tools that may be slower or require more resources.

3. Availability: The Buckwalter analyzer is freely available and can be easily integrated into different software applications and platforms.

4. Transliteration scheme: The Buckwalter analyzer uses a simple and easy-to-use transliteration scheme that maps Arabic characters to ASCII characters, making it easy to input and process Arabic text in computer systems.

However, there are some limitations to the Buckwalter analyzer. For example:

1. Limited language coverage: The Buckwalter analyzer is designed specifically for Modern Standard Arabic and may not work as well for other dialects or styles of Arabic.

2. Morphological complexity: While the Buckwalter analyzer is accurate in identifying the root form and morphological features of Arabic words, it may not capture all of the nuances and complexities of Arabic morphology.

3. Lack of semantic information: The Buckwalter analyzer focuses primarily on morphology and may not provide much information about the meaning or semantics of Arabic text.

In a way, the Buckwalter Arabic Morphological Analyzer is a powerful tool for natural language processing of Arabic text, particularly for tasks such as lemmatization, part-of-speech tagging, and named entity recognition. Its accuracy, speed, and ease of use make it a popular choice among researchers and developers working with Arabic text. The Buckwalter Arabic Morphological Analyzer can be integrated into different software applications and platforms in several ways:

1. Command line interface: The analyzer can be run from the command line, using a simple interface that takes Arabic text as input and produces morphological analysis as output. This makes it easy to incorporate the analyzer into custom scripts and applications.

2. API: The analyzer is also available as an API (Application Programming Interface), which allows developers to integrate it into web applications and other software platforms. The API provides a

simple interface for sending Arabic text to the analyzer and receiving the output as structured data.

3. Plugins: The Buckwalter analyzer is available as a plugin for several popular NLP tools, such as NLTK, GATE, and UIMA. This makes it easy to use the analyzer within these tools and to take advantage of their other features and functionalities.

4. Libraries: The Buckwalter analyzer is also available as a library for several programming languages, such as Python, Java, and C++. This allows developers to integrate the analyzer into their own software applications and to take advantage of the features and functionalities provided by the library.

Thus, the Buckwalter Arabic Morphological Analyzer is a versatile tool that can be integrated into a wide range of software applications and platforms, making it a popular choice among researchers and developers working with Arabic text. The Buckwalter Arabic Morphological Analyzer can be used with several popular Natural Language Processing (NLP) tools, including:

1. **NLTK** (Natural Language Toolkit): NLTK is a widely used platform for building Python programs to work with human language data. The Buckwalter analyzer can be used as a plugin for NLTK, allowing users to take advantage of its features and functionalities.

2. **GATE** (General Architecture for Text Engineering): GATE is an open-source software platform for language processing tasks, such as information extraction, text mining, and sentiment analysis. The Buckwalter analyzer can be used as a plugin for GATE, allowing users to analyze Arabic text within the GATE environment.

3. **UIMA** (Unstructured Information Management Architecture): UIMA is a software architecture for building NLP systems, developed by IBM. The Buckwalter analyzer can be used as a component within UIMA, allowing users to analyze Arabic text within the UIMA framework (Verspoor & Baumgartner, 2013).

4. **Stanford NLP**: The Stanford NLP toolkit is a suite of open-source software tools for natural language processing, developed by Stanford University. The Buckwalter analyzer can be integrated with the Stanford NLP toolkit, allowing users to analyze Arabic text within the Stanford environment.

5. **Apache OpenNLP**: Apache OpenNLP is an open-source toolkit for natural language processing, developed by the Apache Software

Foundation. The Buckwalter analyzer can be used as a plugin for OpenNLP, allowing users to analyze Arabic text within the OpenNLP framework.

These are just a few examples of the many NLP tools that the Buckwalter Arabic Morphological Analyzer can be used with. The flexibility and versatility of the analyzer make it a popular choice among researchers and developers working with Arabic text.

11.I-Arabic: A Suggested Platform

I-Arabic is a language learning platform that focuses on computationally preserving Modern Arabic language and cultural heritage. The platform offers a range of courses and resources for learners at different levels of proficiency, as well as researchers in different fields. Some features of I-Arabic include:

1. Interactive lessons: I-Arabic offers interactive lessons that include videos, audio recordings, and exercises to help learners practice their Arabic language skills.

2. Personalized learning: I-Arabic uses adaptive learning technology to personalize the learning experience for each student, tailoring the content and pacing of the lessons to their individual needs and abilities.

3. Cultural immersion: I-Arabic emphasizes the importance of cultural immersion in language learning, and includes resources on Arabic culture, history, and traditions.

4. Gamification: I-Arabic uses gamification to make language learning fun and engaging, incorporating elements such as badges, leaderboards, and rewards to motivate learners to practice regularly.

5. Live tutoring: I-Arabic offers live tutoring sessions with experienced Arabic language instructors, providing learners with personalized feedback and support as they progress through the course.

Moreover, I-Arabic is a comprehensive and engaging language learning platform that can be a great option for non-native speakers looking to learn Arabic. Its range of courses, personalized learning features, and emphasis on cultural immersion make it a valuable resource for learners at all levels of proficiency.

12.Concluding Remarks

In conclusion, computational attempts to modern Arabic not only have made significant strides in recent years, but also have faced significant challenges. The unique features of the Arabic language, including its non-linear script and complex morphology, have posed substantial challenges for natural language processing and machine learning applications. However, significant progress has been made in areas such as machine translation, speech recognition, and natural language generation.

Despite these advances, computational approaches to Arabic language processing remain limited in their scope and effectiveness. The development of accurate and comprehensive language models for Arabic requires substantial resources and expertise, and the lack of standardized data sets and evaluation methods poses significant challenges for researchers in the field. Additionally, the diglossic nature of Arabic, with its distinct dialects and regional variations, further complicates computational approaches to the language.

Nonetheless, the benefits of computational approaches to Arabic language processing are significant. These approaches have the potential to improve access to Arabic language content, facilitate cross-cultural communication, and support language learning and education. By continuing to invest in research and development in this area, it may be possible to overcome the challenges posed by the diglossic nature of Arabic and develop effective computational tools and resources for the language. As the field continues to evolve, it is important to remain aware of the challenges and limitations of computational approaches, while also recognizing the potential benefits of these approaches for the Arabic-speaking world and beyond.

References:

1. Abbas, M., & Smaili, K. (2005). 'Comparison of topic identification methods for the Arabic language'. In *Proceedings of International Conference on Recent Advances in Natural Language Processing*, RANLP pp 14-17.
2. Ahmad, A. A. S., Hammo, B., & Yagi, S. (2017). 'Construction of an English-Arabic Political Parallel Corpus' *New Trends in Information Technology* (NTIT)–2017, 2, 93. pp 157-171.
3. Ahmed, A., Ali, N, Alzubaidi, M. Zaghouani, W. Abd-alrazaq, A., Househ, M. (2022). 'Free and Accessible Arabic Corpora: A Scoping Review', *Computer Methods and Programs in Biomedicine Update*, 100049. Available from https://www.sciencedirect.com/science/article/pii/S2666990022000015 [Accessed 8 February 2023]
4. Al-Ajmi, H. (2004). A new english–arabic parallel text corpus for lexicographic applications. *Lexikos*, 14, 326–330.
5. Alansary, S., & Nagi, M. (2014). 'The international corpus of Arabic: Compilation, analysis and evaluation'. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing* ANLP, pp. 8-17.
6. Alfaifi, A., & Atwell, E. (2016). Comparative evaluation of tools for Arabic corpora search and analysis. *International Journal of Speech Technology*, 192, 347–357.
7. Alghyaline, S. (2023). Arabic Optical Character Recognition: A Review. *Computer Modeling in Engineering & Sciences*, 135(3), 1825-1861
8. Al-Jawfi, R. (2009). Handwriting Arabic character recognition LeNet using neural network. Int. Arab J. Inf. *Technol*., 63, 304–309.
9. Alotaibi, H. M. (2016). 'AEPC: Designing an Arabic/English parallel corpus', *Research in Corpus Linguistics*, pp 1-7.
10. Alrabiah, M., Al-Salman, A., & Atwell, E. S. (2013). 'The design and construction of the 50 million words KSUCCA'. In *Proceedings of WACL'2 Second Workshop on Arabic Corpus Linguistics*, The University of Leeds, pp 5-8.
11. Al-Saif, A., & Markert, K. (2010). 'The Leeds Arabic discourse treebank: Annotating discourse connectives for Arabic' In

Proceedings of the seventh international conference on language resources and evaluation LREC'10). pp 2046-2053.
12. Al-Sulaiti, L., & Atwell, E. S. (2006). The design of a corpus of contemporary Arabic. *International Journal of Corpus Linguistics*, 112, 135–171.
13. Al-Thubaity, A., Khan, M., Al-Mazrua, M., & Al-Mousa, M. (2013). 'New language resources for Arabic: corpus containing more than two million words and a corpus processing tool' In 2013 *International Conference on Asian Language Processing* pp 67-70. IEEE.
14. Atwell, E. (2018). 'Classical and modern Arabic corpora: Genre and language change'. In RJ. Whitt, (ed.), *Diachronic Corpora, Genre, and Language Change. Studies in Corpus Linguistics*, 85, pp 65-91. John Benjamins.
15. Baker, M. (2019). 'Corpus Linguistics and Translation Studies: Implications and applications' In: Kim, K.H., & Zhu, Y. (eds.), *Researching Translation in the Age of Technology and Global Conflict*. (pp 9-24). Routledge.
16. Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. *Target. International Journal of Translation Studies*, 7(2), 223–243.
17. Baker, M., Francis, G., & Tognini-Bonelli, E. (eds.) (1993). *Text and technology: in honour of John Sinclair*. John Benjamins Publishing.
18. Belinkov, Y., Habash, N. Kilgarriff, A., Ordan, N., Roth. R., and Suchomel, V. (2013). ArTenTen: A new, vast corpus for Arabic. Retrieved from: https://www.sketchengine.eu/wp-content/uploads/arTenTen_corpus_for_Arabic_2013.pdf
19. [Accessed February 20 2023].
20. Biel, Ł. (2014). The textual fit of translated EU law: A corpus-based study of deontic modality. *The Translator*, 20(3), 332–355.
21. Boudelaa, S., & Marslen-Wilson, W. D. (2010). Aralex: A lexical database for modern standard Arabic. *Behavior Research Methods*, 422, 481–487.
22. Brierley, C., & El-Farahaty, H. (2019). An interdisciplinary corpus-based analysis of the translation of كرامة karāma, 'dignity' and its collocates in Arabic-English constitutions. *The Journal of Specialised Translation* (JoSTrans), 32, 121–145.

23. Darwish, K. (2014) Arabizi Detection and Conversion to Arabic *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Langauge Processing (ANLP)*, pages 217–224, October 25, 2014, Doha, Qatar. Association for Computational Linguistics
24. Dukes, K., & Atwell, E. (2012). 'LAMP: A multimodal web platform for collaborative linguistic analysis'. In *Proceedings of the Eight International Conference on Language Resources and Evaluation* LREC'12) (pp. 3268-3275). (European Language Resources Association ELRA).
25. Dukes, K., Atwell, E., & Habash, N. (2013). Supervised collaboration for syntactic annotation of quranic Arabic. *Language Resources and Evaluation*, 471, 33–62.
26. El-Farahaty, H., & Elewa, A. (2020). A Corpus-based analysis of deontic modality of obligation in Arabic–English constitutions'. *Estudios De Traducción*, 10, 107–136.
27. El-Haj, M., & Koulali, R. (2013). 'KALIMAT a multipurpose Arabic Corpus'. In *The Second Workshop on Arabic Corpus Linguistics* WACL-2, pp. 22-25.
28. El-Haj, M., Kruschwitz, U., & Fox, C. (2015). Creating language resources for under-resourced languages: Methodologies, and experiments with Arabic. *Language Resources and Evaluation*, 493, 549–580.
29. Elkhafaifi, H. (2021). Language Planning in the Arab World in an Age of Anxiety. In K. Ryding & D. Wilmsen (Eds.), *The Cambridge Handbook of Arabic Linguistics* (Cambridge Handbooks in Language and Linguistics, pp. 32-47). Cambridge: Cambridge University Press. doi:10.1017/9781108277327.003
30. Good, J. & Cysouw, M. (2013). "Languoid, doculect, and glossonym: formalizing the notion 'language'". *Language Documentation & Conservation*. 7: 331–359.
31. Goweder, A., & De Roeck, A. (2001). 'Assessment of a significant Arabic corpus'. In *Arabic NLP Workshop* at ACL/EACL.
32. Habash, N., Zalmout, N., Taji, D., Hoang, H., & Alzate, M. (2017). 'A parallel corpus for evaluating machine translation between Arabic and European languages'. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*: Volume 2, Short Papers: (pp. 235-241).

33. Habash, N. Y. (2010). Introduction to Arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 31, 1–187.
34. Hajlaoui, N., Kolovratnik, D., Väyrynen, J., Steinberger, R., & Varga, D. (2014). 'Dcep-digital corpus of the european parliament'. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC'14) (pp. 3164-3171).
35. Khwaileh, T., Mustafawi, E., Herbert, R., & Howard, D. (2018). Gulf Arabic nouns and verbs: A standardised set of 319 object pictures and 141 action pictures, with predictors of naming latencies. *Behavior Research Methods*, 506, 2408–2425.
36. Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). The sketch engine. In *Proceedings of the 11th EURALEX International Congress*, (pp. 105-116).
37. Kruger, A. (2004). 'Corpus-based translation research comes to Africa'. *Language Matters: Studies in the Languages of Southern Africa*, 35, 1–5.
38. McCarthy, M. &. O'Keeffe, A. (2012). 'Analysing Spoken Corpora'. In C. A. Chappelle (eds.). *The Encyclopedia of Applied Linguistics*. DOI: https://doi.org/10.1002/9781405198431. Online at: http://onlinelibrary.wiley.com/doi/10.1002/9781405198431.wbeal0028/full.
39. McEnery, T., Hardie, A., & Younis, N. (2019). 'Introducing Arabic Corpus Linguistics'. In T. McEnery, A. Hardie, & N. Younis (eds.), *Arabic Corpus Linguistics*, (pp. 1–16). Edinburgh University Press. Available from http://www.jstor.org/stable/10.3366/j.ctvcwndq8.4 [Accessed February 25 2022]
40. Parkinson, D. B. (2012). *ArabiCorpus*. Online. Available from: https://arabicorpus.byu.edu/ [Accessed February 20 2023]
41. Rühlemann, C. (2019). *Corpus linguistics for pragmatics: A GUIDE FOR RESEARCH*. Routledge.
42. Salhi, H. (2013). Investigating the complementary polysemy and the Arabic translations of the noun destruction' in EAPCOUNT. *Meta: Journal des Traducteurs/Meta: Translators Journal*, 58(1), 227–246.
43. Sharaf, A., Atwell, E. S., Dukes, K., Sawalha, M., Al-Saif, A., Sharoff, S. & Roberts, A. (2010). 'Arabic and Quranic

computational linguistics projects at the University of Leeds' المشاريع الحاسوبية على اللغة العربية والقرآن بجامعة ليدز./Almashārīc Al-hāsūbiyyah cala Al-lughah Al-crabiyyah fī jāmicat Leeds'. In *Proceedings of the workshop of Increasing Arabic Contents on the Web, Organised by Arab League Educational, Cultural and Scientific Organization* (ALECSO).

44. Sharaf, A. B., & Atwell, E. (2012a). 'QurAna: Corpus of the Quran annotated with pronominal anaphora'. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, (LREC'12): (pp. 130-137).
45. Sharaf, A. B., & Atwell, E. (2012b). 'QurSim: A corpus for evaluation of relatedness in short texts'. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*. (LREC'12): (pp. 2295-2302).
46. Sharoff, S. (2006). Open-source corpora: Using the net to fish for linguistic data. *International Journal of Corpus Linguistics*, 114, 435–462.
47. Verspoor, K., Baumgartner, W.A. (2013). Unstructured Information Management Architecture (UIMA). In: Dubitzky, W., Wolkenhauer, O., Cho, KH., Yokota, H. (eds.) *Encyclopedia of Systems Biology*. Springer, New York, NY. https://doi.org/10.1007/978-1-4419-9863-7_183
48. Zaghouani, W. (2017). 'Critical survey of the freely available Arabic corpora', Available from https://arxiv.org/abs/1702.07835 [Accessed November 12 2022]
49. Zaki, M. (2020). 'Corpus-based language teaching and learning: Applications and implications', *International Journal of Applied Linguistics*, 6 October 4th Quarter/Autumn.
50. Zaki, M., Wilmsen, D., & Abdulrahim, D. (2021). 'The Utility of Arabic Corpus Linguistics', *The Cambridge Handbook of Arabic Linguistics*, pp 473-503.
51. Zaki, M. (2021). 'Corpora and translation teaching in the Arab world'. In Said M. Shiyab (eds.), *Research into Translation and Training in Arab Academic Institutions*, (pp. 21-40).
52. Zeroual, I., & Lakhouaja, A. (2018). 'Arabic corpus linguistics: major progress, but still a long way to go. In Shaalan, K.,

Hassanien, A. E., & Tolba, F. (eds.), *Intelligent Natural Language Processing: Trends and Applications*:(pp. 613-636).

53. Ziemski, M., Junczys-Dowmunt, M., & Pouliquen, B. (2016). 'The United Nations parallel corpus, In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, (LREC'16) (pp. 3530-3534).